Basic course on regression modelling

Mervi Eerola

Turku Center of Statistics

24-25.9 and 1.10.2015

Outline of the course

1. Basic principles of statistical inference
2. Linear regression: the basics
3. Linear regression: modifying the model
4. Logistic regression

Gelman, A. & Hill, J. (2007): Data analysis using regression and multilevel/hierarchical models. Analytical methods for social research. Cambridge University Press.

Linear regression: centering

- Interactions allow the predictive effect change in subgroups and may therefore be important for model interpretation
- Large main effects often have interactions and should be checked
- As seen already, the intercept may not always have a meaningful interpretation ($=$ values of predictors are 0).
- Centering each predictor variable around its mean or around some meaningful reference point helps interpretation
- This, and other transformations of the predictors, which fit better to the data, are discussed later

Statistical inference in linear regression: notation

- *Units:* individuals, schools, cities, families etc. (In multilevel models: pupils in schools, members in families etc)
- *Outcome variable* $y_i, i = 1, ..., n$ for n units
- *Predictor variables* X, usually a matrix of p variables where x_{ip} is the pth predictor variable for unit i
- *Errors* ϵ_i of the model: the random (not explained) part of the model
- Error terms are assumed to follow the distribution $N(0, \sigma^2)$ where the parameter σ^2 represents the variability with which the outcomes deviate from their predictions based on the model

Statistical inference in linear regression: the model

- The linear regression model for unit i is written as

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i,$$

  where x_{i1} is a constant term, defined to be 1 for each unit
- Example (Child's test score): x_{i2}=`mother.hs`, x_{i3}=`mother.IQ` and $x_{i4} = x_{i2} \cdot x_{i3}$ the interaction term
- Another way to write the linear regression model is to define

$$y_i \sim N(X_i \beta, \sigma^2), \text{for all } i = 1, ...n$$

- Solving the model corresponds to finding the 'best' estimates (least squares or maximum likelihood) to the regression coefficients β and the residual variance σ^2

Regression estimates and their uncertainty

- The least squares estimates $\hat{\beta}$ of the regression parameters minimize the sum of squared errors $\sum_{i=1}^{n}(y_i - X_i\hat{\beta})^2$, the error of prediction
- Estimation of β introduces uncertainty which is presented in the *standard errors* of the coefficients
- Coefficient estimates with roughly 2 standard errors of $\hat{\beta}$ are considered reliable (cf. confidence limits)
- High correlation between the predictors X (multicollinearity) increases the standard errors of the coefficients
- Residuals $r_i = y_i - X_i\hat{\beta}$ are differences between observed data and model prediction

Unexplained variance

- Residual standard deviation $\hat{\sigma} = \sqrt{\sum_{i=1}^{n} r_i^2/(n-p)}$ summarizes the *unexplained* part of the model and the scale of the residuals
- $n - p$ is the degrees of freedom (number of free data minus the number of estimated parameters)
- Example (Child's test score): The residual standard deviation is $\hat{\sigma} = 18$ points which is the accuracy of our model: it is the average distance of each test score observation from its model prediction (cf. range of scores $[20, 140]$ points)

Explained variance

- The *explained* part of the variance in data (coefficient of variation) is the fraction explained by the predictors

$$R^2 = 1 - \hat{\sigma}^2/\hat{\sigma}_y^2$$

  where $\hat{\sigma}_y$ is the standard deviation of the data (total variation)
- Example (Child's test score): The squared multiple correlation $R^2 = 22\%$ is not very high, indicating that child's ability cannot be predicted from mother's IQ and education only
- Note. Statistically significant coefficients $\hat{\beta}$ are roughly 2sd's away from zero but even insignificant predictors should be included in the model if there is a theoretical reason for it

Assumptions and diagnostics

The following assumptions in linear regression are ordered by importance:

1. *Validity of the model*. Is your outcome variable really measuring your problem? Are all relevant predictors included? Does your sample allow for the conclusions you wish make?
2. *Additivity and linearity.* The deterministic part is a linear function of the predictors. If not, try other forms of the predictors (e.g log(x), x^2)
3. *Independence of errors.* The deviations from the regression line are assumed independent.
4. *Equal variance of errors*. If in doubt, use weighted least squares.
5. *Normality of errors*. Least important. In small samples, use t-distribution instead of normal.

3. Linear transformations of the data

1. Linear transformations
2. Centering and standardizing
3. Correlation and regression to the mean
4. Logarithmic transformations
5. Regression models for prediction

Linear transformations of the data

- Sometimes we need to transform the variables to achieve additivity and linearity, and sometimes to aid interpretation of the model as a description of the underlying mechanism
- Linear transformations do not change the fit of the model or model predictions
- Scaling of predictors changes the size of regression coefficients: recall that the coefficient represents the average change in the outcome due to one unit change in the value of the predictor itself
- Example (Earnings and height): height in inches:
  $\text{earnings} = -6100 + 1300 \cdot \text{height} + \text{error}$
  height in millimeters:
  $\text{earnings} = -6100 + 51 \cdot \text{height} + \text{error}$

Standardization: z-scores

- Example (Earnings and height): Regression of earnings on height can be modelled in inches or centimeters but the interpretation of a 1sd change in either is the same
- Standardization using z-scores: Replace `height` with `z.height` by defining

$$z.height = (\text{height} - \text{mean}(\text{height}))/\text{sd}(\text{height})$$

- Interpretation of coefficients: a one standard deviation change in the predictor x produces a change of size β in y
- Interpretation of intercept: mean of y when all predictors x are at their mean values

Centering and standardizing

- The coefficients of models with interactions are meaningless if the reference takes the value 0 (cf. IQ=0 or heigth=0). The reference group does then not exist.
- Centering the predictor simplifies the interpretation:

```
c.mother.hs=mother.hs - mean(mother.hs)
c.mother.iq=mother.iq - mean(mother.iq)
```

- The residual standard deviation, R^2 and the regression coefficient of interaction do not change, but the main effects and intercept *do* change and have now a meaningful interpretation
- Note. The reference point can also be some theoretically meaningful value. What could they be in this example and how do the interpretations change?

Scaling problem

- Binary or discrete predictors have always larger coefficients than continuous predictors: the coefficient of `mother.hs` is much larger than that of `mother.IQ`
- Scaling (standardizing) the predictors by $2sd_x$ rather than by sd_x maintains rough comparability between coefficients of binary and continuous predictors

```
 z.mother.hs = mother.hs - mean(mother.hs)/2*sd(mother.hs)
z.mother.iq = mother.iq - mean(mother.iq)/2*sd(mother.iq)
```

  and they can be interpreted on a common scale
- If there are no interactions, the same effect is achieved by multiplying the regression coefficients of the *original* predictors by $\hat{\beta} \cdot 2sd_x$

Correlation and regression

- Standardize a single predictor regression model

$$y = \alpha + \beta x + \text{error}$$

  by centering and scaling

$$\frac{y - \bar{y}}{\text{sd}_y} = \beta^* \frac{x - \bar{x}}{\text{sd}_x} + \text{error}$$

  so that `z.y`$=\beta^*$`z.x`+error
- Then β^* is the *correlation* between y and x and the intercept $\alpha = \bar{y} + \beta\bar{x} = 0$
- Thus the value of the regression slope of two standardized variables must be between -1 and 1
- In general, the regression slope in a model with one predictor is $\beta = \rho(\sigma_y/\sigma_x)$ where ρ is the correlation between y and x and σ_y and σ_x are the standard deviations

Regression to the mean

- When y and x and standardized, the slope is always less than 1: when the predictor x is 1sd away from the mean, the predicted value of y is *less* than 1sd above the mean
- This phenomenon of the linear regression model is called *regression to the mean* and and holds for any pair of variables that are not perfectly correlated
- Examples: predicting the height of sons (daughters) with the height of fathers (mothers) would result in less extreme deviation from the mean in sons (daughters)
- The actual observed value of the heights of sons and daughters can of course differ from the *predicted* value by the size of the residual term

Logarithmic transformation

- A key assumption of simple regression is additivity and linearity
- If the values of the outcome are all-positive, it is often better to model the *logarithms* of the outcome
- We then consider *relative* rather than absolute changes in y
- To interpret the results, back-transform to the original scale by exponentiating the coefficients
- Note. The model is *multiplicative* rather than additive

$$log(y_i) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

because

$$y_i = exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i) = e^{\alpha} \cdot e^{\beta_1 x_{i1}} \cdot e^{\beta_2 x_{i2}} \cdot e^{\epsilon_i}$$

- The exponential terms are positive so the predicted values of y are positive

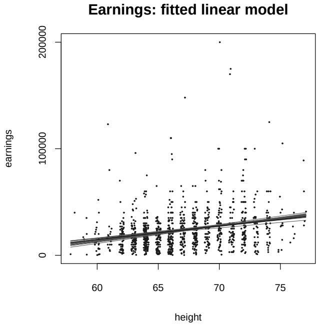
Example: Height and earnings

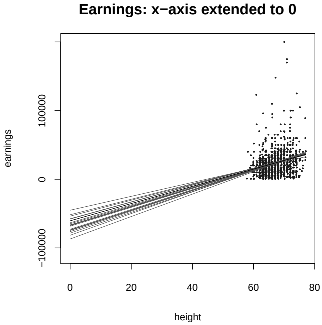
- Consider the predictive effect of height on earnings
- `Earnings` is all-positive; take the logarithms
- Note: individuals with `earnings`$= 0$ must be excluded

```
earn.logmodel.1 <- lm(log.earn ~ height)
> display(earn.logmodel.1)
            coef.est  coef.se
(Intercept)  5.78      0.45
height       0.06      0.01
---
n = 1192, k = 2
residual sd = 0.89, R-Squared = 0.06
```

- The predicted effect of 1 inch increase in height is 6% on earnings because $exp(0.06) \sim 1.06$ (relative change)
- Does this apply to both sexes or does the fact that men are taller explain the result?

Example: Height and earnings





Example: Height and earnings

- Do taller people earn more, on average, than shorter people of the *same* sex? Include `male` as a predictor.

```
> earn.logmodel.2 <- lm(log.earn ~ height + male)
> display(earn.logmodel.2)
            coef.est coef.se
(Intercept) 8.15     0.60
height      0.02     0.01
male        0.42     0.07
---
n = 1192, k = 3
residual sd = 0.88, R-Squared = 0.09
```

- The effect of 1 inch increase is now less, 2%, but the comparison between sexes is $exp(0.42) = 1.52$; being male increases earnings by 52% when differences in heigth are controlled

Statistical significance of the results

- A simple check of significance: if the coefficient of the predictor x is larger than $2sd_x$, it is statistically significant: the estimate is included in the 95% confidence interval
- Example (Height and earnings): the 2% effect of height is still statistically significant and on \$50 000 earnings it corresponds to a \$1000 increase
- The overall effect of the predictors can again be judged by the multiple correlation coefficient R^2 and the estimated residual standard deviation $\hat{\sigma}$
- R^2 measures the explained part (due to height and sex) of the model and σ the unexplained part

Residual standard deviation and R^2

- Example (Height and earnings): The residual standard deviation $\hat{\sigma} = 0.88$ indicates that 68% of log earnings will be within 0.88 of the predicted or average value of log-earnings
- On the original scale this corresponds to a factor of $exp(0.88) = 2.4$
- For example, a 70 inch person has predicted earnings $8.153 + 0.021 \cdot 70 = 9.623$, with a predictive standard deviation on average 0.88
- With 68% confidence, this person has log-earnings 9.623 ± 0.88, on the original scale in the interval $[exp(8.74), exp(10.50)] = [6000, 36000]$ dollars.

Residual standard deviation and R^2

- This very wide range tells that the model does not explain very much of the variation in earnings; this is also reflected in $R^2 = 0.09$ which implies that only 9% of the variation is due to height and sex, although both are statistically significant
- Adding the interaction between sex and height does not change the model; the estimate is positive but not statistically significant. For prediction purposes it could however be included in the model.
- Note that the interpretation of intercept is again problematic: `height`=0
- Centering and scaling (standardization) help; the comparison is then against a person of average height

Example: Height and earnings

- Is height more important for men or for women? Consider interaction 'heigth:male'
- Standardize height because `height`= 0 meaningless in interaction

```
> z.height <- (height - mean(height))/sd(height)
> earn.logmodel.4 <- lm(log.earn ~ z.height + male + z.height:male)
> display(earn.logmodel.4)
                coef.est coef.se
(Intercept)     9.53     0.05
z.height        0.07     0.05
male            0.42     0.07
z.height:male   0.03     0.07
---
n = 1192, k = 4
residual sd = 0.88, R-Squared = 0.09
```

- The interaction is not significant but has an expected sign: Increase in height is more beneficial for men than for women

Centered interaction model: interpretation of coefficients

- `Intercept`: predicted log earnings for a woman of average height 66.9 inches
- `z.height`: predicted difference in log earnings corresponding to a 1sd difference in height for a woman (`male`=0); estimated predictive difference for a 3.8 inch increase in height is 7% for women
- `male`: predictive difference in log earnings between men and women when `z.height` = 0; The ratio is $exp(0.42) = 1.52$ or 52% more for a 66.9 inch man than a woman of the same size
- `z.height:male`: difference in slope of the predictive differences for height among men and women; a 3.8 inch increase in height corresponds to 3% more increase in earnings for men than for women, and in total $3\% + 7\% = 10\%$
- Note: This model compares men and women of equal height!

Log-log transformation

- What if log transformation is applied to `height` also?
- Then the regression coefficient is interpreted as the expected *proportional change* in earnings per *proportional change* in height
- In economics, the coefficients in a log-log model are called 'elasticities'
- In general, if the range of values (ratio: high/low values) is close to 1, log transformation does not make much difference
- However, the interpretation of coefficients may be easier to understand on the log scale; proportional increase in earnings per inch or proportional increase in height

Example: Earnigns and heigth: Log-log model

- Proportional changes in outcome and predictor: log-earn and log-height

```
> log.height <- log(height)
> earn.logmodel.5 <- lm(log.earn ~ log.height + male)
> display(earn.logmodel.5)
            coef.est coef.se
(Intercept) 3.62     2.60
log.height  1.41     0.62
male        0.42     0.07
---
n = 1192, k = 3
residual sd = 0.88, R-Squared = 0.09
```

- For a 1% change in height there is a predicted 1.41% change in earnings given sex